

A mixture model approach to infer land-use influence on point referenced water quality

Adrien Ickowicz · Jessica Ford · Keith Hayes

the date of receipt and acceptance should be inserted later

Abstract The assessment of water quality across space and time is of considerable interest for both agricultural and public health reasons. The standard method to assess the water quality of a catchment, or a group of catchments, usually involves collecting point measurements of water quality and other additional information such as the date and time of measurements, rainfall amounts, the land-use and soil-type of the catchment and the elevation. Some of this auxiliary information will be point data, measured at the exact location, whereas other such as land-use will be areal data often in a compositional format. Two problems arise if analysts try to incorporate this information into a statistical model in order to predict (for example) the influence of land-use on water quality. First is the spatial change of support problem that arises when using areal data to predict outcomes at point locations. Secondly, the physical process driving water quality is not compositional, rather it is the observation process that provides compositional data. In this paper we present an approach that accounts for these two issues by using a latent variable to identify the land-use that most likely influences water quality. This latent variable is used in a spatial mixture model to help estimate the influence of land-use on water quality. We demonstrate the potential of this approach with data from a water quality research study in the Mount Lofty range, in South Australia.

1 Introduction

Accounting for spatial association is becoming an increasingly important topic in ecological data analysis. Spatially referenced data are typically observed either at points in space (point-referenced or simply point data) or over areal units such as counties or zip codes (block data). The change of support problem (COSP; see [Gelfand et al., 2001](#)) "is concerned with inference about the values of the variable at points or blocks different from those at which it has been observed". Change of support problems of interest include predicting the dependent variable for a

CSIRO, Hobart, 7004 TAS, Australia

particular area given values of explanatory variables measured at points at different locations within that area (see for example [Zhu et al. \(2003\)](#); [Cressie and Wikle \(2011\)](#)) or the reverse: predicting some point-estimation given area-wide observations of explanatory variables. It is the later problem that usually occurs in the context of water quality monitoring (see [Beck, 1987](#)). Typically the analyst is presented with point-referenced observations (of water quality parameters such as concentrations of metals or nutrients, that may be used to ensure compliance with potable water quality standards. A natural way to approach this problem statistically is to develop a model of the spatio-temporal dependencies, where the spatial dependence is captured by the covariance matrix of the error term, and the temporal dependence is captured either through seasonality parameters (see [Lindstrom et al., 2011, 2013](#)) or via an auto-regressive process (AR, see [Bakar and Sahu, 2013](#)). Typically, however, other parameters such as land-use type will be known (or suspected) to have an effect on water quality, and should therefore be incorporated into the model. At this point two complications may occur. The first arises because land-use information is often recorded as a compositional measure (see [Aitchison, 2003](#); [van den Boogaart and Tolosana-Delgado, 2013](#)). For example, land-use is often recorded as a vector showing the proportion of the catchment under each of the use-types. The second complication is the COSP. This arises because land use is measured over an area that includes the point measurements of water quality. Various methods exist for predicting the effect of land use on water quality at catchment scales. One common method is the Soil and Water Assessment Tool (SWAT, see [Arnold et al., 1995](#)) is a sophisticated, continuously distributed simulation model. It operates on a daily time step and is designed to predict the effect of land use, land management practices, and climate change on the quality and quantity of surface and ground water (see <http://swat.tamu.edu/>). SWAT assumes an in-depth knowledge of the mechanistic processes that govern water quality and quantity within a watershed. It requires the analyst to quantify the parameters that govern the rates of these processes (such as surface runoff, percolation, Evapotranspiration, etc.) This level of understanding and information, however, is not always available (often because the cost of acquiring it at large scales is prohibitive)). In view of this alternative statistical methods have also been developed, the simplest of which is referred to as the 'lumped' approach (see [Strayer et al., 2003](#); [King et al., 2005](#)). The lumped approach treats compositional observations as covariates in a linear model, sometimes relying on transformations of the explanatory and/or response variables (see [Buck et al., 2004](#)) or the derivation of response indices or metrics (see [Shen et al., 2014](#)) to work within the confines of the linear modelling framework. The lumped approach also assumes that "each portion of the catchment has equal influence" on the water quality ([Peterson et al., 2011](#)). This approach, however, does not address the COSP and this may lead to incorrect estimates, particularly if the areas of the catchment partition are large. For example, if a land-use type such as "forested highlands" occupies a large proportion of a catchment, but never occurs close to a water quality monitoring site (often in high-order river stretches), then it would be reasonable to anticipate that it would have only a small effect on the water quality at the monitoring site), whereas under the lumped approach its influence

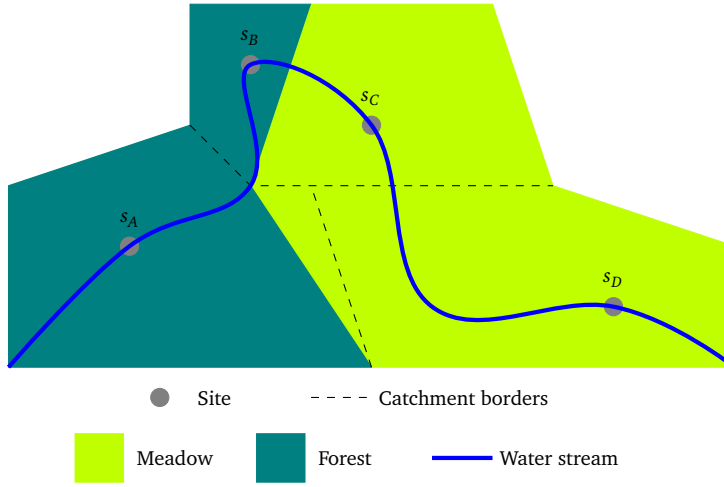
would be proportional to its area in the catchment.

One way to avoid this problem with the lumped approach is to incorporate a "distance to the water" measurement in the statistical model, and [Peterson et al. \(2011\)](#) presents a list of techniques to achieve this. The basic idea is to modify the effect of land-use area by incorporating a weight proportional to the inverse distance between the land-use type and the point of measurement. This weighting may also allow for additional considerations, for example it may incorporate the effects of flow accumulation and dilution effects through a river network ([Hunsaker and Levine, 1995](#)) or the outputs of simple hydrological models ([Burcher, 2009](#)). These approaches require knowledge of the land-use distribution across the whole catchment, in order to build the distance matrix, and are sensitive to decisions about how to measure distance between areal units and points. [Tong and Chen \(2002\)](#) takes a different approach to this problem by first constructing a non-parametric test to exclude land-use types that are not correlated with the mean of the water quality parameters within a particular hydrological unit (but still ignoring the COSP), and then using the remaining land-use covariates in a process-based simulation model. This has the advantage of reducing the model dimension, and also helps to solve the issue of non-influential land-use categories mentioned earlier. When testing the correlation between land-use and water quality in the first stage, however, the spatial-temporal structure of the problem is ignored. This could cause misleading results. For example, a land-use category may be uncorrelated with the mean of a water quality parameter, and be therefore discarded in the initial stage of the analysis, but it may have a seasonal influence that would be missed in the subsequent modelling.

Another technique specifically designed to deal with the COSP is area-to-point kriging, described in [Kyriakidis \(2004\)](#) and [Yoo and Kyriakidis \(2006\)](#), and applied in [Bonyah et al. \(2013\)](#). Area-to-point kriging is an interpolation technique, that provides an interpolated compositional value at each location of measurement. For a given location, each composition value is a weighted average of the surrounding lattices compositions. Each weight carries both the dimension of the lattice and its distance to the location of interest. While this limits the influence of a distant/small lattice composition on the location, the covariates still belong to the composition space when the response variable is not, by essence, subject to compositional covariates. In this paper we develop a mixture modelling approach to estimate the effect of land use on water quality, using the latent variable to identify which land-use type influences water quality at monitoring site. Mixture models have been studied for a long time, and their strengths and weaknesses are well known (see [McLachlan and Peel, 2000](#), for a comprehensive review). Mixture models with compositional data, however, are not common, although some examples can be found within the existing literature ([Ongaro et al., 2008](#); [Meinicke et al., 2011](#)). Mixture models offer some advantages in this context, for example, accounting for any spatial dependence between the latent variables in several ways, including:

- A mixture of expert with spatial random effects ([Neelon et al., 2014](#));

Fig. 1 Measurements in the same catchment are not subject to the same land-use influence. s_B and s_C are specific examples.



- A mixture with a discrete or continuous (Markov random field - Potts or Gaussian process model) prior on the latent variable (Woolrich et al., 2005).

In this article, we demonstrate how a modified version of the latter approach provides an alternative to the methods described previously. The main advantage of our approach is that it deals with the COSP while also capturing the spatial and seasonal effects of land-use type on water quality. The disadvantages of this approach are that it limits the land-use influence to one type per site, for any given water quality parameter, and does not include an effect for the area of the land use type.

The article is organized as follows. In section 3, we present the model used to infer the spatio-temporal effects of land use types on water quality. In section 4 we present two estimation approaches. The first uses the Expectation-Maximisation (EM, see Dempster et al., 1977) algorithm. The second adopts a Bayesian hierarchical approach. In section 5 we compare the results obtained with our approach to the results obtained with an existing model drawn from the literature. We apply our method to a dataset from a water monitoring program from sub-catchment of the Mount Lofty Range in South Australia (see Ford et al., 2015, for details of the program).

2 Data and problem

The Mount Lofty Ranges (MLR) are important in South Australia (SA) because they provide significant water resources to a range of stakeholders, including agricultural landholders, secondary industries and potable water suppliers and consumers. A draft Water Allocation Plan (WAP) for the MLR was released in 2010-2011. In addition to the WAP the SA government identified the need for improved water quality in the MLR catchments through the Water Quality Improvement and the Water for Good programs [Ford et al. \(2015\)](#). As part of these programs, water quality was monitored in 18 sites over a period of 14 years (1998 – 2012). Measurements were not collected every day due to resource limitations, but the frequency of observations was at least once per week. Various parameters including basic physico-chemical variables such as turbidity, dissolved oxygen, EC, pH and temperature as well as more investigation-specific parameters such as dissolved organic carbon (DOC) [Varcoe et al. \(2010\)](#) and nutrient and pesticide concentrations [Cox et al. \(2012\)](#) were recorded. These programs also recorded the land-use types (as a composition variable) in the catchments that contribute to the water being measured at each of the 18 monitoring sites. The land-use type is recorded within a hierarchical classification scheme. The first (coarsest) level of the hierarchy distinguishes 6 broad categories of land use. The second level splits this classification into 32 more detailed categories, and the third (finest) level further sub-divides the land use into 85 categories. In this analysis we use the 8 most well represented (and interpretable) land use categories from the second level primarily because the number of measurement sites was so small (to do otherwise would result in a very sparse design matrix and in over-parametrization issue).

Assessing the influence of land-use on total nitrogen with these data proves to be difficult for the following reasons:

1. We don't know what the land-use type is in the immediate vicinity of the measurement site - we only know the composition of land-use types in the catchment that the station is located in;
2. We don't know the distance along the stream network between measurement sites. We could calculate the Euclidean distance but this is not an optimal metric;

Our analysis aims to identify the effect of land-use type on water quality, and thereby provide a mechanism to predict water quality at sites in catchments that have not been monitored despite the aforementioned limitations. This analysis was designed to form part of a wider assessment of the risks of exceeding water quality parameters across the MLR, particularly under low flow conditions [Ford et al. \(2015\)](#). Here we use the concentration of total nitrogen measured at 16 sites from 2008 to 2013 to demonstrate application of the statistical model.

3 Modeling

3.1 Framework

Our data consists of geo-referenced, time stamped observations of Nitrogen concentration at 16 sites in the MLR. This immediately suggests the need for a spatio-temporal analysis. Szpiro et al. (2010) (and then Sampson et al., 2011; Lindstrom et al., 2011) propose the following general approach model for this type of data:

$$y(s, t) = \sum_j f_j(t) \beta_j(s) + \nu(s, t) \quad (1)$$

where $y(s, t)$ denotes the observation at site s and time t , $\beta_j \sim MVN(\tilde{\beta}_j, \Sigma_\beta)$ captures the (spatially varying) effect of site-specific covariates and $\nu \sim N(0, \Sigma_\nu)$ is a space-time residual field. The temporal variation in the data is decomposed into three basis functions $f_j, j = 1 \dots 3$ representing long term trend, seasonal effects and random variation respectively (as detailed in Cleveland and Cleveland, 1990) and $\tilde{\beta}_j$ captures the mean effect of site specific covariates on each of these components of temporal variation. In our analysis the relevant site-specific covariates are the land-use types in the catchment(s) that influence the water quality at a monitoring station.. In the general model, this 'land-level' information is represented by Z_j , such that:

$$\beta_j(s) = \alpha_j Z_j(s) + \eta(s) \quad (2)$$

where $\eta(s) \sim N(0, \Sigma_\eta)$. It is clear from Equation 2 that :

- the temporal basis functions are the same at each location. The effect of the land-level covariates (e.g. land-use type) is modelled through α_j hence these covariates can only influence the amplitude of the basis function;
- the spatial structure of the influence of the land-level covariates relies on the spatial correlation structure of β_j . Hence, it can be described using a stationary universal Kriging;
- a common, and important simplification, stipulates that Σ_η is diagonal. This approach treats the water quality at each monitoring site as spatially independent, influenced only by (for example) the land-use type in its immediate catchment, and ignores the influence that other catchments might have by virtue of the fact that monitoring sites may be connected by a stream network.

In this paper we present an approach that aims at improving this model in the following ways:

1. Land use types are typically recorded as a proportion of the surrounding catchment. This means that for each measurement site, a set of land-use types is observed, with the presumption that the probability the land-use types influences water quality is related in some way to the proportion of the catchment that they occupy. . In these circumstances, we believe that a mixture model is more appropriate than transforming the land-use observation using common transformations (see Aitchison, 2003, for a list of possible transformations.).

2. We relax the single temporal basis function assumption by allowing each mixture component to have its own set of temporal basis functions. For instance, given a land-use type k , measurements at site i and time t can be model using the following equation:

$$y_i(t)|k = \sum_j f_j(k, t) \beta_j(i) + \nu \quad (3)$$

3. The spatial structure still relies on the land-use type, but we now use a neighbouring structure, of the type used in image analysis because we believe that modelling the spatial structure through the Euclidean distance between measurement sites is not optimal in our problem. Our interpretation is that the spatial continuity of land usage across multiple catchments is the key factor leading the Total nitrogen measurements characteristics. This translate into a neighbouring structure for the spatial model because of the lattice form of the catchments.

3.2 Incorporating land-use as a latent variable

The general modelling framework we adopt is a point process model $\hat{\mathcal{A}}^{\mathcal{S}}$ i.e. it assumes that the land-use predictors in Z are observed at each of the locations s_i . Land-use information, however, is most often defined over areas often with a significant size. This is a typical spatial misalignment problem, referred to as change of support problem in the introduction.

The land-use information collated in the MLR is presented as a compositional observations of land use types in the immediate catchment of the monitoring site, that is,

$$z_k(S_i) = \frac{1}{|S_i|} \int_{u \in S_i} z_k(u) du \quad (4)$$

where $|S_i|$ is the area for sub-catchment i , and $z_k(u)$ an indicator function for the presence of land-use k at the point location u (in S_i). Ideally, we would build our model on the knowledge of $z_k(s)$, where s is the location of the monitoring station. For example, if we look at Figure 1, $z_F(s_A) = 1$, and $z_F(S_A) = 1$ for land use type $k = \text{'Forest'}$. Which is good for the model, as this avoids the spatial misalignment issue. However, if we look at the second location, $z_F(s_B) = 1$ for land use type 'Forest', but $z_F(S_B)$ is only 0.3 for this land use type. The majority of land-use in area S_B is 'Meadow', and this could result in model estimates of $z_F(s_B) = 0$, leading to inconsistency and bias in the resultant estimators.

What we propose is a mixture model approach. The rationale is the following. Given a single sample location, the land-use at this location has the biggest influence on the water quality measurements. This means that the measured water quality has a value "linked" to the land-use type. Then for multiple sites, the resulting measurements are sampled¹ from a mixture distribution. The number of

¹ *sampled* is used here because we don't know the land-use type at the locations of measurement.

distributions in the mixture is then equal to the number of land-use types in the entire area.

3.3 Spatio-temporal structure of the model

As stated in the previous sections, this problem is fully spatio-temporal and possible correlations have to be integrated in the model. We list below the solutions chosen to account for these correlations.

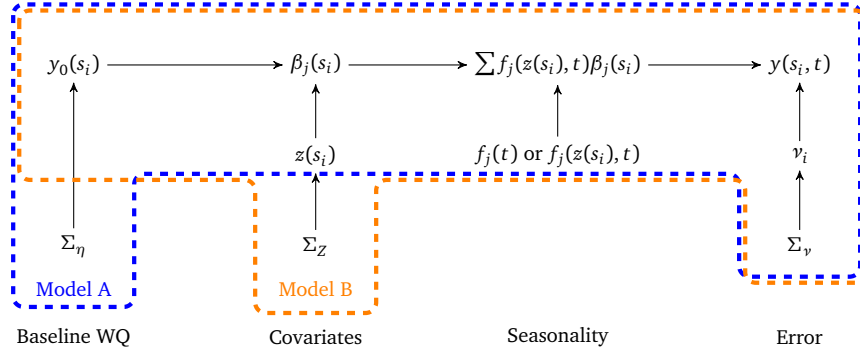
3.3.1 Temporal dependency

The model presented by Szpiro et al. (2010) supports the idea that the temporal variation remains globally identical over the spatial domain, making it possible to fit a smooth temporal function prior to fitting the full model. The spatio-temporal residuals ν are then assumed to be independent in both space and time. This assumption, however, is very strong. We therefore develop an alternative modelling approach that allows for seasonality over the spatial domain, that can be affected by the land-use type at each measurement location, both in terms of amplitude, phase and shape. Note that the original model assumes that this influence affects amplitude alone. Our new model introduces temporal basis functions for each land-use type. These basis functions are calculated using a seasonal-trend (STL) decomposition approach introduced by Cleveland and Cleveland (1990). In this approach the time series of observations are decomposed into trend and seasonality components using a Loess smoother. In our model, the time series for each land-use is extracted from the time series of each site using the latent variable. Then, for a given land-use type, the STL decomposition is applied to each site time-series associated to that land-use through the latent variable.

3.3.2 Spatial dependency

It is possible to make multiple assumptions regarding the spatial structure. The most common approach is to assume that neighbouring sites are correlated with a correlation value that is a function of the Euclidean distance between sites. Recent works (Peterson et al., 2007) suggests that this model is not optimal when we dealing with data on a stream network, as the euclidean distance between sites is less relevant than the distance measured along the network, hence Peterson et al. (2007) suggest the use of stream distance rather than euclidean distance. These two approaches limit the correlation structure to the measurements. We believe that some of the spatial dependency observed in the data come from the spatial structure of the land-uses. The solution presented in this paper is to model the spatial dependency through the latent variable of the mixture, making sure that neighbouring sites (in the sense of common border) are more likely to have identical land-uses. In the EM technique, this can be achieved by adding a penalization term. In the Gibbs sampling technique we model the latent variable of the mixture using a Potts model (also called 2-d Ising model).

Fig. 2 Schematic figure of the model structure. The spatial correlation can be included at different levels of the model, leading to different likelihood and estimation tools. Model A include the spatial correlation at the baseline level (y_0), implying natural spatial correlation without the effect of land-use. This is the model most often used in the literature. Model B, which we use here, introduces the spatial correlation at the covariate level (land-use). Merging the two models can be considered, but one has to carefully monitor for over-fitting by doing so.



3.4 Final model

With the changes and assumptions described previously, the model equation becomes:

$$y(s, t) = \sum_j f_j(z(s), t) \beta_j(s) + v(s, t) \quad (5)$$

where the temporal basis function f is now land-use (location) dependent, and β can be understood as a random effect parameter for each location-temporal basis pair. $z(s)$ is the latent indicator for which land-use is associated to site s .

4 Estimation

The model described in Eq. 1 is very simple to express in the likelihood format, as all the components are Gaussian. This leads to a linear log-likelihood, for which maximization is performed quite straightforwardly (see Lindstrom et al., 2013, for details about the maximization procedure, including some simplification tricks). The introduction of the mixture model on top of the spatial dependency (see Eq. 5) requires us to use an EM algorithm or a dedicated Bayesian hierarchical model. The following sections provide a description of the two possible estimation solutions.

4.1 Pre-requisite

There are two elements that are not estimated by the model, but need to be incorporated in order to perform the estimation: the temporal basis; and the neighbouring structure.

Temporal basis functions Preliminary temporal basis functions $\bar{f}_{ij}(t)$ are first estimated at each site using a Loess seasonal and trend decomposition (Cleveland and Cleveland, 1990). The basis function for each land-use is then calculated from these site specific decomposition using a linear equation,

$$f_j(k, t) = (\sum_i 1_{(Z_i=k)} \bar{f}_{ij}) / \sum_i 1_{(Z_i=k)} \quad (6)$$

for function j , and land-use k . That equation also implies that in the estimation procedure, each time the latent variables are updated, the temporal basis functions are also updated.

Neighbouring structure In order to compute the spatial dependency of the data, we need to identify the spatial structure. In this paper, we assume that the latent variables are carrying all the spatial dependency, by virtue of their neighbourhood relationship. This structure is achieved through the construction of a Voronoi lattice, where the cell centers are the monitoring sites.

4.2 The EM approach

The EM algorithm was initially developed to achieve the estimation task in models with incomplete data (Dempster et al., 1977). As stated in this initial paper, the algorithm is "broadly applicable" and has since been used for mixture models estimation, around the idea that missing data also means latent non-observed (or observable) variable.

Given the observations $(y_i)_{i \leq N}$, and an unobserved latent variable $\delta \in \mathcal{D}$, the EM algorithm aims at maximizing the augmented log-likelihood

$$\ell_N(\theta) = \sum_i \log h_\theta(y_i) \text{ where } h_\theta(y) = \int_{\mathcal{D}} f_\theta(y, \delta) d\delta$$

Because δ is unobserved, it is difficult to evaluate $\ell_N(\theta)$. Instead, the EM maximize the expected (over δ) complete-data likelihood given y and a previously calculated θ .

$$\begin{aligned} Q(\theta; \theta') &= E_\delta[\log p_\theta(y, \delta) | \theta', y] \\ &= \int_{\mathcal{D}} \log p_\theta(y, \delta) p(\delta | \theta', y) d\delta \end{aligned} \quad (7)$$

which can usually be approximated by

$$Q_N(\theta; \theta') = \frac{1}{N} \sum_i \int_{\mathcal{D}} p_{\theta}(\delta_i | y_i) \log p_{\theta'}(y_i, \delta_i) d\delta_i \quad (8)$$

because of the independence assumption between the observations. This simplification step is important because the integral becomes one-dimensional, allowing for easier calculation of its value. In order to achieve the maximization, the EM proceeds in two steps during each iteration,

$$\left| \begin{array}{ll} \text{(E)} & \text{Compute } p_{\theta}(\delta | y_i) \\ \text{(M)} & \text{Set } \theta^{(c+1)} = \underset{\theta}{\operatorname{argmax}} Q_N(\theta, \theta^{(c)}) \end{array} \right. \quad (9)$$

We refer to [Bilmes \(1998\)](#) for a "gentle tutorial" of the EM algorithm.

If we choose Model A (see Figure 2), using covariance matrix Σ_{η} , we cannot "jump" from Eq. 7 to Eq. 8 for two reasons: first, because the complete observations are not independent, even conditionally; second, because the latent variable is not straightforward to define. We have a hierarchical set of latent variables that need to be considered: β is the higher level one; for Z we want to estimate the land-use type. A technical solution would be to use Monte-Carlo EM, where the latent state variables are sampled. Although technically valid, this technique may require a huge number of samples because of the potential dimension of the latent state.

Another important problem is that \mathcal{D} can be huge, depending on the number of land-use predictors and sites. For instance, with p land-use types and n sites, the number of element to sum up in Eq. 7 is p^n (instead of $p \times n$ with the classical EM). It is also common knowledge that the EM algorithm relies heavily on its initialization to achieve the global convergence. Poor initialization can lead to a local maxima only, even more when the dimension of the parameter space increases (see [Wu, 1983](#); [Archambeau et al., 2003](#); [Naim and Gildea, 2012](#)). Different strategies have been proposed to overcome this problem, ([Biernacki et al., 2001](#); [Dicintio, 2012](#); [Baudry and Celeux, 2015](#)), however in the end it is most important to make the initialization as close as possible to the true solution.

On the other hand, by choosing Model B, and assuming that the spatial structure is just as well defined in the land-use space, we can use the Neighborhood EM algorithm, first introduced by [Ambroise et al. \(1997\)](#); [Ambroise and Govaert \(1998\)](#). This algorithm allows to introduce a constraint on the mixture variable in order to account for spatial consistency.

In order to take the spatial dependence of objects into account, they suggest considering partitions which are optimal according to a penalized Hathaway criterion. The term of penalization should favour homogeneous classes. Spatial relationships can be summarized in different ways. In their articles, neighborhood was favored through a boolean matrix,

$$v_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are neighbours} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Then, defining $c_{ik} = \frac{p(\delta_i=k)f(y_i|\theta_k)}{f(y_i|\theta_k)}$ \mathbf{v} is turned into a penalized term,

$$G(c) = \sum_i \sum_j \sum_k c_{ik} c_{jk} v_{ij} \quad (11)$$

and added to Q leading to the following functional to optimize,

$$U(\theta, \theta^{(c)}) = Q(\theta, \theta^{(c)}) + \beta G(c) \quad (12)$$

The overall principle remains the same, with two steps, one for expectation, one for maximisation. [Ambroise and Govaert \(1998\)](#) proposed an optimisation method based on the fixed point approach to achieve the E-step, which is slightly changed because of the penalization term.

4.3 The Bayesian hierarchical model

In the Bayesian framework, latent state models are referred to as Bayesian hierarchical models. In their general formulation, they display three levels:

$$\pi(\theta_1, \theta_2 | y) \propto \underbrace{\pi(y|\theta_1)}_{data} \underbrace{\pi(\theta_1|\theta_2)}_{process} \underbrace{\pi(\theta_2)}_{prior} \quad (13)$$

The data level refers to the likelihood of the observations given the parameters at the process level. The process level refers to the latent process captured by a spatio-temporal model for the data level parameters. In our context, what we gained with this approach is the latent spatial model, needed because the available covariates (land-use) are inadequate to capture the effects (see [Cooley and Sain, 2010](#); [Cooley, 2013](#); [Lehmann et al., 2013](#), for extreme precipitation example).

From the modelling section, we remember the special form of our modelling equation. The main issue we are trying to solve lies in the right hand side of Eq. 5: $z(s)$ is an indicator latent variable, where levels are the different land-use types. In this process, we make the following assumptions:

- only one land-use can have an influence on one station;
- the influence of the land-use is independent of the area of the land-use (although this assumption is not really mandatory, it makes the model simpler. Otherwise, an offset variable can be used to account for the influence of the area of the land);

It is important to note that the first assumption is different to the assumption that only one land-use type affects a catchment. If multiple stations are located in one catchment, many land-use effects can be observed and inferred using this model. The spatial dependency structure of Z is given by modelling it as a hidden Potts model, and in particular as a Gibbs random field. There is a wealth of literature in applied statistics on these methods (see [Cressie, 2015](#); [Rue, 2005](#); [Green and](#)

Richardson, 2002). In a Gibbs model, the probability density function of Z can be written:

$$f(z) = \frac{1}{Z} \exp\left\{-\sum_{c \in \mathcal{C}} U_c(z)\right\} \quad (14)$$

where \mathcal{C} is the neighbourhood of c , and U is a potential function. Here, we define,

$$f(z|\delta) = \frac{1}{Z_\delta} \exp\{\delta^T S(z)\} \quad (15)$$

where $S(z) = \sum_{c' \in \mathcal{C}} 1_{z_c=z_{c'}}$ is the number of neighbours of c that belong to the same mixture. The posterior distribution is defined over the parameters $\theta = \{\nu_k, \beta_k, \delta\}$. Outputs from the model also include the latent variable z and the temporal basis function $f_j(k, \cdot)$. A very convenient way to sample from the posterior distribution is to use a Gibbs sampler. In our model, we assume $\Sigma_{\nu,k} = \nu_k Id$, and we have a Gibbs sampler that is almost explicit,

- $\nu_k \sim IG(\frac{N}{2} + a, \frac{\sum (y_i - \mu_{z_{i,k},k})^2}{2} + b)$, where $\mu_{z_{i,k},k}$ is the predicted value for y_i using the estimated parameters;
- $\mu_{z_{i,k},k} \sim N(\bar{y}_{j,k}, \nu_k/n_{j,k})$ and β_k through a linear transformation of $\mu_{z_{i,k},k}$;
- δ is updated using the scheme of Murray et al. (2006);
- $z_{i,k} \sim \mathcal{M}(1, w_{i,1,k}, \dots, w_{i,K,k})$ with

$$w_{i,j,k} = \frac{\exp[-\frac{1}{2}(\frac{y_i - \mu_{j,k}}{\nu_k})^2 + \delta \sum_{c \in \mathcal{C}_i} 1_{z_{c,k}=j}]}{\sum_j \exp[-\frac{1}{2}(\frac{y_i - \mu_{j,k}}{\nu_k})^2 + \delta \sum_{c \in \mathcal{C}_i} 1_{z_{c,k}=j}]} \quad (16)$$

The main difficulty lies in updating δ . However, quick mixing can be achieved using the scheme of Murray et al. (2006) and the Swendsen-Wang algorithm to simulate from the Potts model. Additional details can be found in Murray et al. (2006); Barbu and Zhu (2007); Everitt (2012); Cucala and Marin (2013).

5 Results

The results of three models are presented in Table 1: one mixture model, one mixture model with spatial dependency and the CLR model. In the table, we present the maximum a posteriori (MAP) estimates for the land-use types, for the different temporal basis function: in our model, three basis function are used, constant, trend and seasonal; in the CLR model two temporal basis functions are used only, to match the degrees of freedom of the models. On the bottom side of the table we display the sum of square errors (SSE, using the MAP) for the prediction power of each model. The three presented models adjust differently to the data. The mixtures approaches have a better fit than the CLR model. Results show that the mixture and the spatial mixture provide almost identical results. This suggests that the estimation of the latent variables demonstrates a natural spatial correlation, which didn't need to be enforced through the spatial modelling. This spatial consistency is also demonstrated in Figure 4.

Table 1 MAP estimates for the three fitted models. It has been ordered from the lowest baseline value, to the highest, following the mixture models estimates. We observe that the order and the amplitude of the estimates is not the same between the mixture models and the compositional-log-ratio (CLR) transform model.

	<i>Dependent variable: Nitrate concentration</i>		
	<i>Spatial mixture model</i>	<i>Mixture model</i>	<i>CLR transform model</i>
Managed resource protection	-4.35	-4.35	-0.73
Nature Conservation	-3.38	-3.38	-0.75
Grazing modified pastures	-3.10	-3.11	-0.49
Plantation forestry	-1.86	-1.89	-
Services, Transport and Comm.	-1.54	-1.50	-0.09
Residential	-0.98	-0.98	-0.22
Irrigated perennial horticulture	0.47	0.47	0.02
SSE	104.62	103.91	121.81

5.1 Estimated land-use influence on the water quality

There are two main outputs from the model that can be analysed in order to understand the influence of land-use on the water quality. The baseline values and the temporal basis functions provide different indications for the ecological expert. The baseline value indicates a de-trended, de-seasonalised level of concentration, which can be used as a summary of the sites water quality. The temporal basis functions indicate the temporal variations, both in trend and seasonality. From Table 1, we observe a strong consistency between the mixture models, while the MAPs estimates for the CLR transform model demonstrate less amplitude and a different order². Figure 3 shows an example of estimated trend and seasonality for the land-uses "nature conservation" and "irrigated perennial horticulture". These figures indicate that:

- "Nature conservation" has a bigger amplitude in seasonality than "irrigated perennial horticulture";
- The amplitude for both land-uses appears to be decreasing;
- Their seasonality appears to have the same phase for both land-uses;
- "Nature conservation" has a trend that indicates a longer seasonality. This is potentially due to external factors and requires further investigation;
- "Irrigated perennial horticulture" shows a decreasing trend.

The different interpretations of the temporal basis function plots can be linked to conditions in the field in order to provide more insight into what is happening (or

² for mathematical constraint reasons the Plantation forestry estimate had to be ignored - All the values are measured and computed on the log-scale.

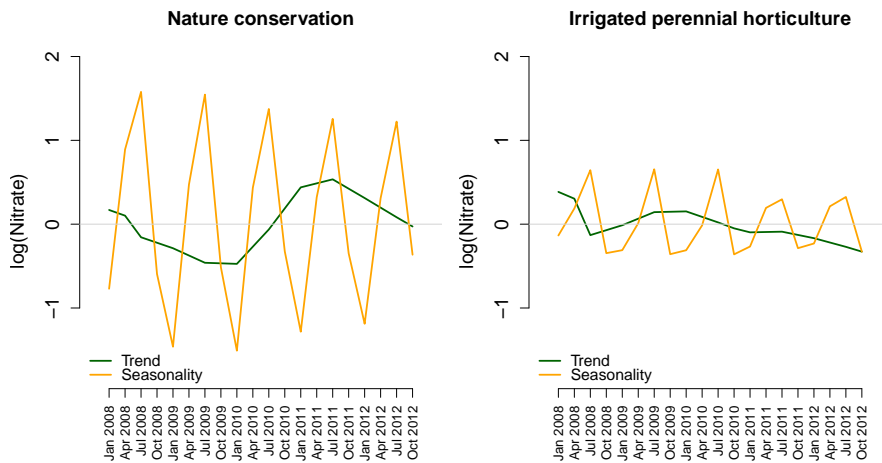


Fig. 3 Example of plot of the (normalized) temporal basis functions for the land-uses "nature conservation" and "irrigated perennial horticulture". The orange line represents the seasonality, the green line represents the trend.

will happen) to water quality in this area. This also can allow for the identification of change points (observable through the trend plot) and for the prediction of water quality in changing conditions.

5.2 Measure of uncertainty

In Eq. 5, the random variable v is representing the uncertainty of the observations. This is usually linked to unknown predictors and measurement errors. However, when the model is over-fitting the data, the uncertainty can be under estimated. This situation is encountered by the CLR model when an additional temporal basis function is added. The resulting SSE and σ_v^2 are both estimated to be equal to 0. With the proposed model the over-fitting risk can be monitored by looking at the smoothness of the seasonality and trend curves. Figure 3 shows regular curves, a result consistently overruling the over-fitting risk. An additional feature of our model is interesting for measuring uncertainty. The mixture approach allows us to decided whether a single measurement error is modelled, or if each component of the mixture has a different measurement error. This additional flexibility allows us to identify sites / land-uses where additional sampling effort are required in order to improve the monitoring.

5.3 Predicting the land-use type location

Another interesting outcome of the model is the map of estimated latent land-use type. For each station, the model can predict the most likely land-use type. Because

of the nature of the model, the latent state is estimated for each Voronoi cell in each sub-catchment, allowing us to display a map of most likely land-use types. We display in Figure 4 a map of the sub-catchments, coloured by their estimated latent state and ordered by their baseline values. We notice the strong spatial consistency of the map, with the higher concentration of nitrate located in the same area.

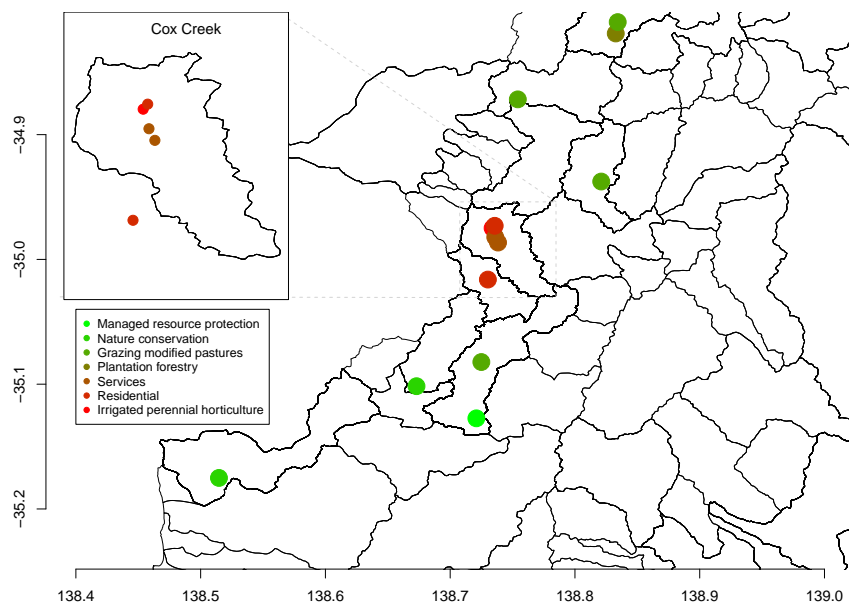


Fig. 4 Map of some of the sub-catchment in the Mount Lofty region. In the legend, items have been ordered according to Nitrate concentration: lower Nitrate concentration values are shaded green (eg MRP); higher Nitrate concentration are shaded red (eg IPH). Higher nitrate measurements are observed near the Cox Creek sub-catchment, with land-use being related to human activities (residential, services, transport and communication, irrigated perennial horticulture).

6 Discussion

In this article, we presented a model that takes into account the land-use influence on the water quality parameters when the supporting data is limited and misaligned. We presented a mixture model approach that allows the latent variables to be spatially correlated. A Bayesian algorithm was proposed to fit the model, and the method was applied to real data. We believe this model delivers two important messages. First, compositional data do not imply compositional modelling, even

if the latter is easily implementable. Spatial misalignment is an important issue that should be considered carefully. Additionally, the presented model provides a helpful insight on the land-use influence when the information about it is limited, particularly in precision. The outputs of the model provide an idea about the land-use and sites association, and associates the variation of the measurements to the most likely land-use influence. Being able to identify these is of prime importance in particular for understanding the real impact of land-use on water quality. Some improvements are possible. First, we limit the number of influence per site to one, as our aim is to identify the main land-use influence per site. This may prove a bit restrictive as some sites may be exposed to more than one significant land-use influence. A simple example for that is when the water source is a river, and the two opposite banks have two different land-use types. Our approach can be generalized to more than one influence, however this would increase the computational load significantly, even more so if we consider that different sites may have a different number of influences (and hence adding to the burden the need to estimate the number of influences for each site). The latent variable approach forces the Markov chain to run longer, primarily due to the Metropolis step within the Gibbs sampler. It is important to monitor the behaviour of the samples of the posterior distribution in order to ensure convergence of the chain. Another potential improvement would be the use of an automatic selection tool for the latent space dimension. In this article, we choose which and how many land-use type constitute the latent space. This approach, based on expert knowledge, can be completed by a variable selection approach (of the type LASSO - [Raftery and Dean \(2004\)](#); [Städler et al. \(2010\)](#), or slope heuristic - [Baudry et al. \(2012\)](#)).

Acknowledgements

This study was supported by funding from the Goyder Institute for Water Research for the project "Mt Lofty Ranges Water Allocation and Planning" led by Associate Professor Jim Cox. The authors would like to thank SA Water, SA EPA, DEWNR, AMLR NRMB and Dr Leon van der Linden for providing water quality data from the Mt Lofty Ranges for the risk assessment. Thank you also to staff from numerous South Australian Government Agencies who provided valuable feedback at presentations given about project outputs.

References

- Aitchison, J., 2003. A concise guide to compositional data analysis. CDA Workshop, Girona .
- Ambroise, C., Dang, V.M., Govaert, G., 1997. Clustering of spatial data by the EM algorithm, in: *geoENV I - Geostatistics for Environmental Applications, Quantitative Geology and Geostatistics* (Vol. 9), pp. 493–504.
- Ambroise, C., Govaert, G., 1998. Convergence of an EM-type algorithm for spatial clustering. *Pattern Recognition Letters* 19, 919–927.

- Archambeau, C., Lee, J.A., Verleysen, M., 2003. On Convergence Problems with the EM for Finite Gaussian Mixtures, in: *Neural Networks*, pp. 99–106.
- Arnold, J.G., Williams, J.R., Maidment, D.R., 1995. Continuous-Time Water and Sediment-Routing Model for Large Basins. *Journal of Hydraulic Engineering* 121, 171–183.
- Bakar, K., Sahu, S., 2013. spTimer: Spatio-Temporal Bayesian Modelling Using R. *Journal of Statistical Software*.
- Barbu, A., Zhu, S.C., 2007. Generalizing Swendsen-Wang for Image Analysis. *Journal of Computational and Graphical Statistics* 16, 877–900.
- Baudry, J.P., Celeux, G., 2015. EM for mixtures - Initialization requires special care. Technical Report. INRIA.
- Baudry, J.P., Maugis, C., Michel, B., 2012. Slope heuristics: Overview and implementation. *Statistics and Computing* 22, 455–470.
- Beck, M.B., 1987. Water quality modeling: A review of the analysis of uncertainty. *Water Resources Research* 23, 1393–1442.
- Biernacki, C., Celeux, G., Govaert, G., 2001. Strategies for Getting the Highest Likelihood in Mixture Models. Technical Report RR-4255. INRIA.
- Bilmes, J.A., 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute* 4, 1–15.
- Bonyah, E., Munyakazi, L., Asong, D., Bashiru, I.I., 2013. Application of area to point Kriging to breast cancer incidence in Ashanti Region of Ghana. *International Journal of Medicine and Medical Sciences* 5, 67–74.
- van den Boogaart, K.G., Tolosana-Delgado, R., 2013. *Analyzing Compositional Data with R*. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Buck, O., Niyogi, D.K., Townsend, C.R., 2004. Scale-dependence of land use effects on water quality of streams in agricultural catchments. *Environmental pollution (Barking, Essex : 1987)* 130, 287–99.
- Burcher, C.L., 2009. Using simplified watershed hydrology to define spatially explicit 'zones of influence'. *Hydrobiologia* 618, 149–160.
- Cleveland, R., Cleveland, W., 1990. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics* 6, 3–73.
- Cooley, D., 2013. Modeling Both Climate and Weather Spatial Effects for Extreme Precipitation. Technical Report. Colorado State University.
- Cooley, D., Sain, S.R., 2010. Spatial hierarchical modeling of precipitation extremes from a regional climate model. *Journal of Agricultural, Biological, and Environmental Statistics* 15, 381–402.
- Cox, J.W., Oliver, D.P., Fleming, N.K., Anderson, J.S., 2012. Off-site transport of nutrients and sediment from three main land-uses in the Mt Lofty Ranges, South Australia. *Agricultural Water Management* 106, 50–59.
- Cressie, N.A.C., 2015. *Statistics for Spatial Data*. Wiley.
- Cressie, N.A.C., Wikle, C., 2011. *Statistics for Spatio-Temporal Data*. Wiley.
- Cucala, L., Marin, J.M., 2013. Bayesian inference on a mixture model with spatial dependence. *Journal Of Computational And Graphical Statistics* 22, 584–597.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series*

- B (Methodological) 39, 1–38.
- Dicintio, S., 2012. Comparing Approaches to Initializing the Expectation-Maximization Algorithm. Ph.D. thesis. University of Guelph.
- Everitt, R.G., 2012. Bayesian Parameter Estimation for Latent Markov Random Fields and Social Networks. *Journal of Computational and Graphical Statistics* 8600, 26.
- Ford, J., Ickowicz, A., Oliver, D., Hayes, K., Kookana, R., 2015. Integrated catchment water planning support for Adelaide Mount Lofty Ranges Water Allocation Planning (GWAP Project) Task 5 : Tiered Water Quality Risk Assessment. Technical Report 15/4. Goyder Institute for Water Research. Adelaide.
- Gelfand, a.E., Zhu, L., Carlin, B.P., 2001. On the change of support problem for spatio-temporal data. *Biostatistics (Oxford, England)* 2, 31–45.
- Green, P.J., Richardson, S., 2002. Hidden Markov Models and Disease Mapping. *Journal of the American Statistical Association* 97, 1055–1070.
- Hunsaker, C.T., Levine, D.A., 1995. Hierarchical Approaches of Water Quality in Rivers Study processes are important in developing. *Sciences-New York* 45, 193–203.
- King, R.S., Baker, M.E., Whigham, D.F., Weller, D.E., Jordan, T.E., Kazyak, P.F., Hurd, M.K., 2005. Spatial Considerations for Linking Watershed Land Cover To Ecological Indicators in Streams. *Ecological Applications* 15, 137–153.
- Kyriakidis, P., 2004. A geostatistical framework for area to point spatial interpolation. *Geographical Analysis* 36, 259–289.
- Lehmann, E.A., Phatak, A., Soltyk, S., Chia, J., Lau, R., Palmer, M., 2013. Bayesian hierarchical modelling of rainfall extremes, in: 20th International Congress on Modelling and Simulation, Adelaide, Australia, 1-6 December 2013, pp. 1–6.
- Lindstrom, J., Szpiro, A., Sampson, P., Sheppard, L., Oron, a., Richards, M., Larson, T., 2011. A flexible spatio-temporal model for air pollution: Allowing for spatio-temporal covariates. *UW Biostatistics Working Paper Series* 370, 1–38.
- Lindstrom, J., Szpiro, A., Sampson, P.D., Bergen, S., Sheppard, L., 2013. SpatioTemporal : An R Package for Spatio-Temporal Modelling of Air-Pollution. CRAN Vignettes .
- McLachlan, G., Peel, D., 2000. *Finite Mixture Models*. John Wiley & Sons, New York.
- Meinicke, P., Aßhauer, K.P., Lingner, T., 2011. Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics* 27, 1618–1624.
- Murray, I., Ghahramani, Z., MacKay, D.J.C., 2006. MCMC for doubly-intractable distributions. *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)* , 359–366.
- Naim, I., Gildea, D., 2012. Convergence of the EM Algorithm for Gaussian Mixtures with Unbalanced Mixing Coefficients. *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* , 1655–1662.
- Neelon, B., Gelfand, A.E., Miranda, M.L., 2014. A multivariate spatial mixture model for areal data: Examining regional differences in standardized test scores. *Journal of the Royal Statistical Society. Series C: Applied Statistics* 63, 737–761.
- Ongaro, A., Migliorati, S., Monti, G.S., 2008. A new distribution on the simplex containing the Dirichlet family. *CoDaWork 2008, the 3rd International Work-*

- shop on Compositional Data Analysis .
- Peterson, E.E., Sheldon, F., Darnell, R., Bunn, S.E., Harch, B.D., 2011. A comparison of spatially explicit landscape representation methods and their relationship to stream condition. *Freshwater Biology* 56, 590–610.
- Peterson, E.E., Theobald, D.M., Ver Hoef, J.M., 2007. Geostatistical modelling on stream networks: Developing valid covariance matrices based on hydrologic distance and stream flow. *Freshwater Biology* 52, 267–279.
- Raftery, A.E., Dean, N., 2004. Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association* 101, 168–178.
- Rue, H., 2005. Gaussian Markov Random Fields: Theory and Applications. *Hand The* 104, 263 p.
- Sampson, P.D., Szpiro, A.a., Sheppard, L., Lindström, J., Kaufman, J.D., 2011. Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmospheric Environment* 45, 6593–6606.
- Shen, Z., Hou, X., Li, W., Aini, G., 2014. Relating landscape characteristics to non-point source pollution in a typical urbanized watershed in the municipality of Beijing. *Landscape and Urban Planning* 123, 96–107.
- Städler, N., Bühlmann, P., van de Geer, S., 2010. 11-Penalization for Mixture Regression Models. *Test* 19, 209–256.
- Strayer, D.L., Beighley, R.E., Thompson, L.C., Brooks, S., Nilsson, C., Pinay, G., Naiman, R.J., 2003. Effects of Land Cover on Stream Ecosystems: Roles of Empirical Models and Scaling Issues. *Ecosystems* 6, 407–423.
- Szpiro, A.A., Sampson, P.D., Sheppard, L., Lumley, T., Adar, S.D., Kaufman, J.D., 2010. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. *Environmetrics* 21, 606–631.
- Tong, S.T., Chen, W., 2002. Modeling the relationship between land use and surface water quality. *Journal of Environmental Management* 66, 377–393.
- Varcoe, J., van Leeuwen, J.A., Chittleborough, D.J., Cox, J.W., Smernik, R.J., Heitz, A., 2010. Changes in water quality following gypsum application to catchment soils of the Mount Lofty Ranges, South Australia. *Organic Geochemistry* 41, 116–123.
- Woolrich, M., Behrens, T., Beckmann, C., Smith, S., 2005. Mixture models with adaptive spatial regularization for segmentation with an application to fMRI data. *Medical Imaging, IEEE Transactions on* 24, 1–11.
- Wu, C., 1983. On the convergence properties of the EM algorithm. *Annals of Statistics* 11, 95–103.
- Yoo, E.H., Kyriakidis, P.C., 2006. Area-to-point Kriging with inequality-type data. *Journal of Geographical Systems* 8, 357–390.
- Zhu, L., Carlin, B., Gelfand, A., 2003. Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta. *Environmetrics* , 1–33.